The Islamic University of Gaza

Civil Engineering Department

Statistics

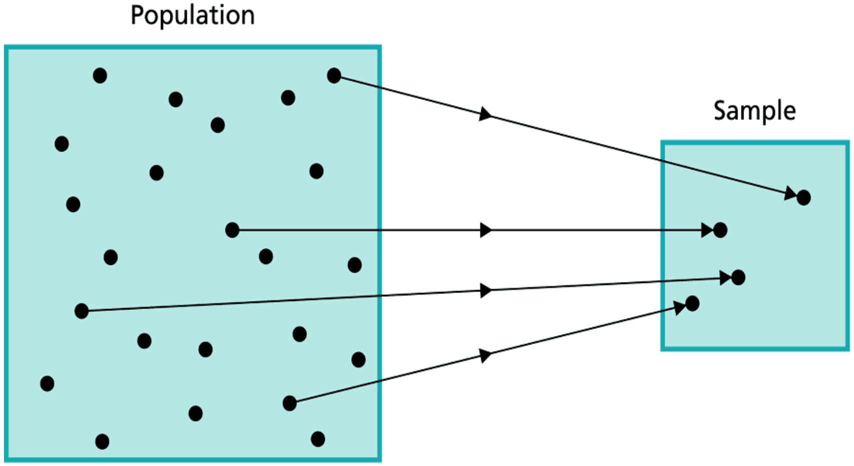
ECIV 2305

# Chapter 6 – Descriptive Statistics

# 6.1 Experimentation

- **Population** consist of all possible observations available for a particular probability distribution.

- **A sample**: is a particular subset of the population that an experiment measures and uses to investigate the unknown probability distribution.



- **A random sample**

  Is one in which the elements of the sample are chosen at random from the population and this procedure is often used to ensure that the sample is **representative** of the population.

The data observations $x_1$, $x_2$, $x_3$, ……, $x_n$ can be grouped into:

1. **Categorical data:**

   mechanical, electrical, misuse

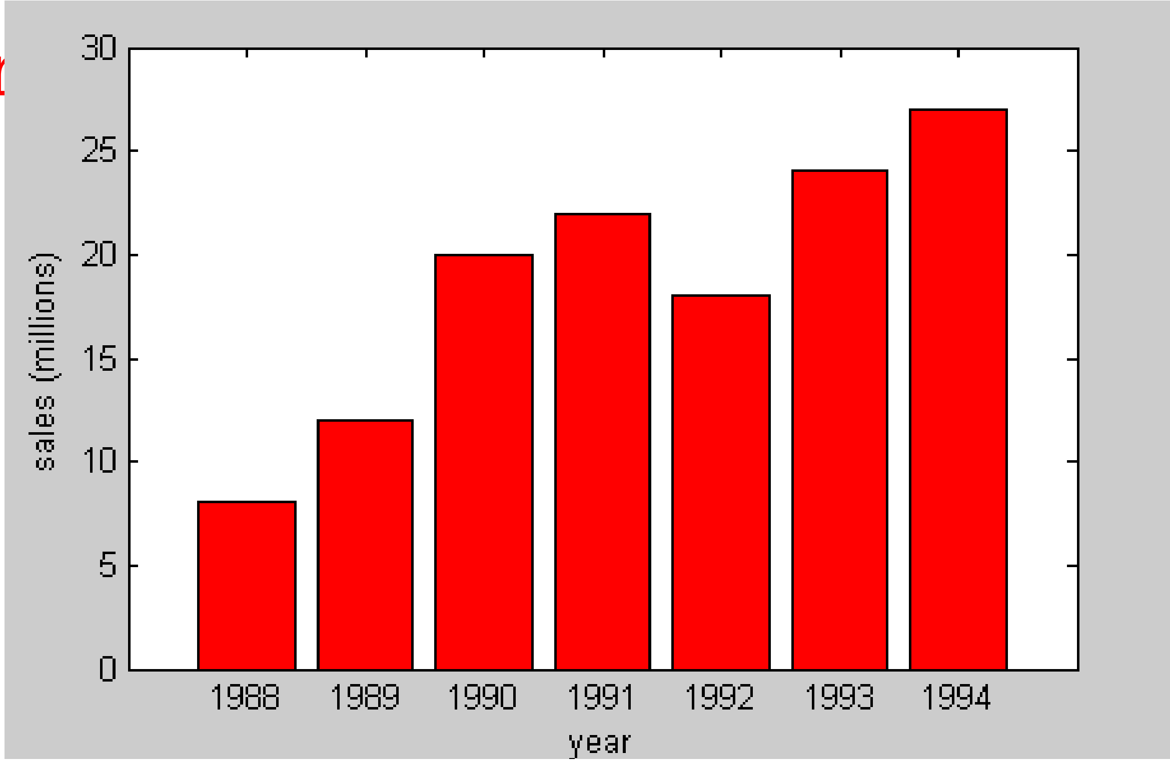
2. **Numerical data** :

   integers or real numbers

# 6.2 Data Presentation

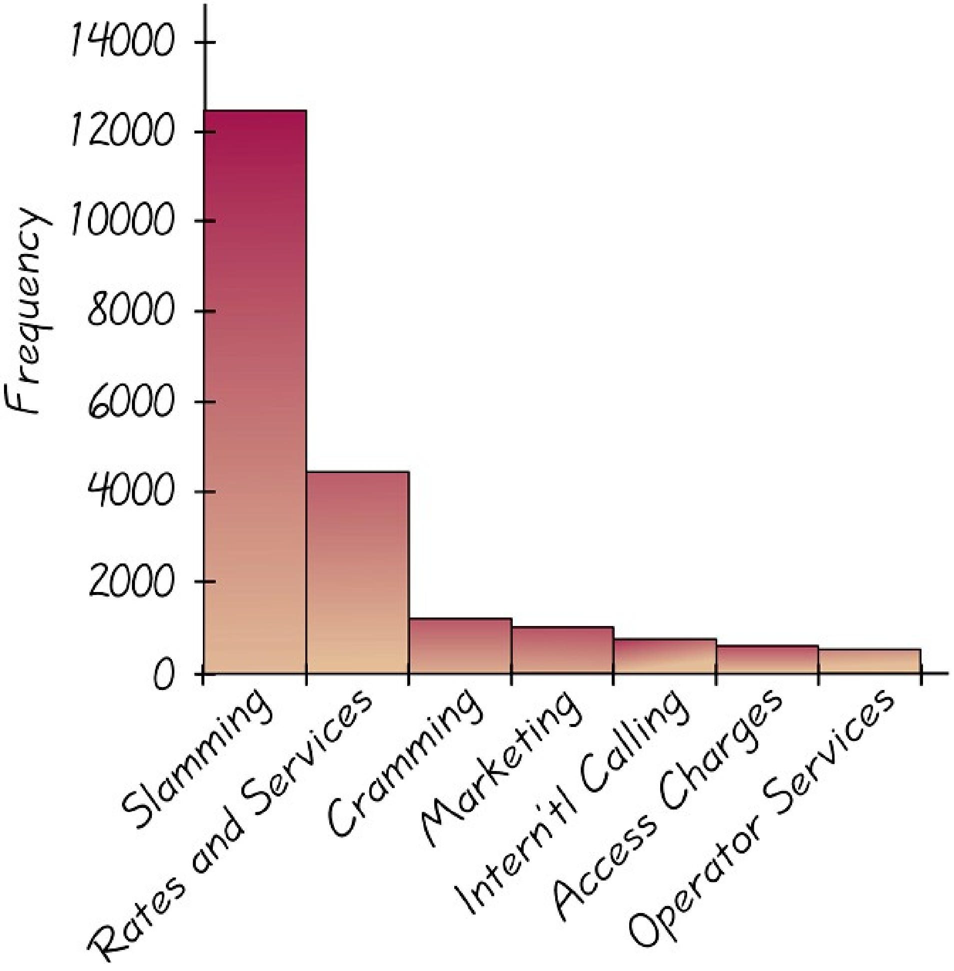
## A. Categorical Data:

**1. Bar Char**



# 2. Pareto Chart

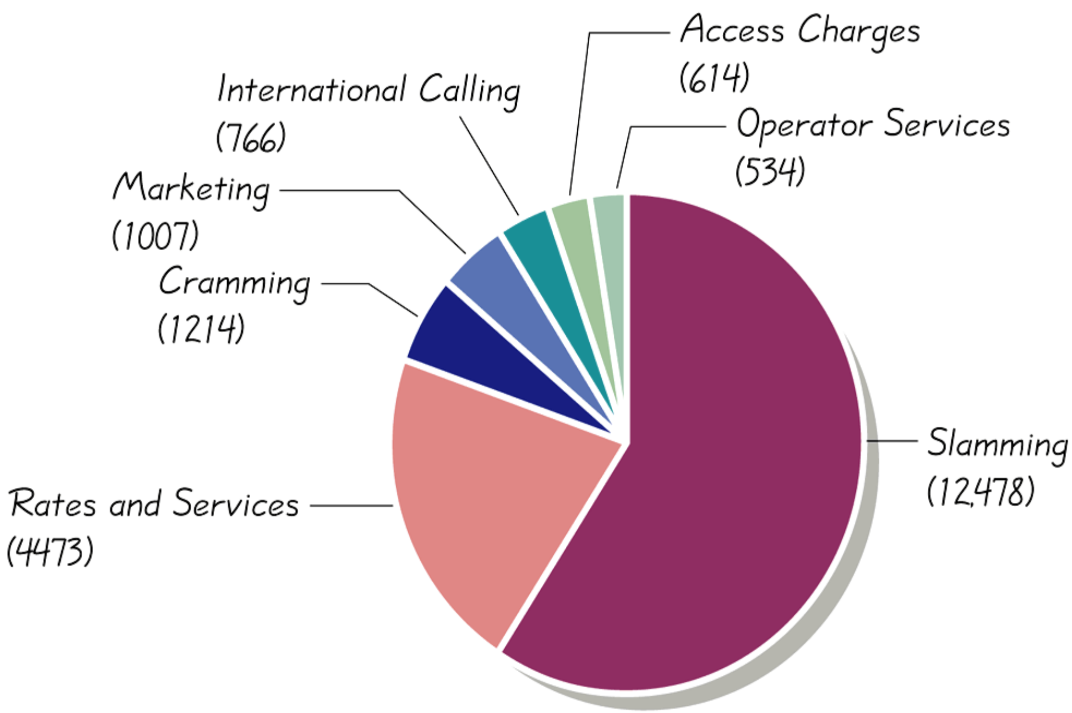
The categories are arranged in order of decreasing frequency.



# 3. Pie Charts

A graph depicting data as slices of a pie

- n = data set of n observations
- r = observations of specific category.
- Slice = (r/n)*360



# 6.2.3 Histograms

- Histograms are used to present numerical data rather than categorical data.

**Example:**

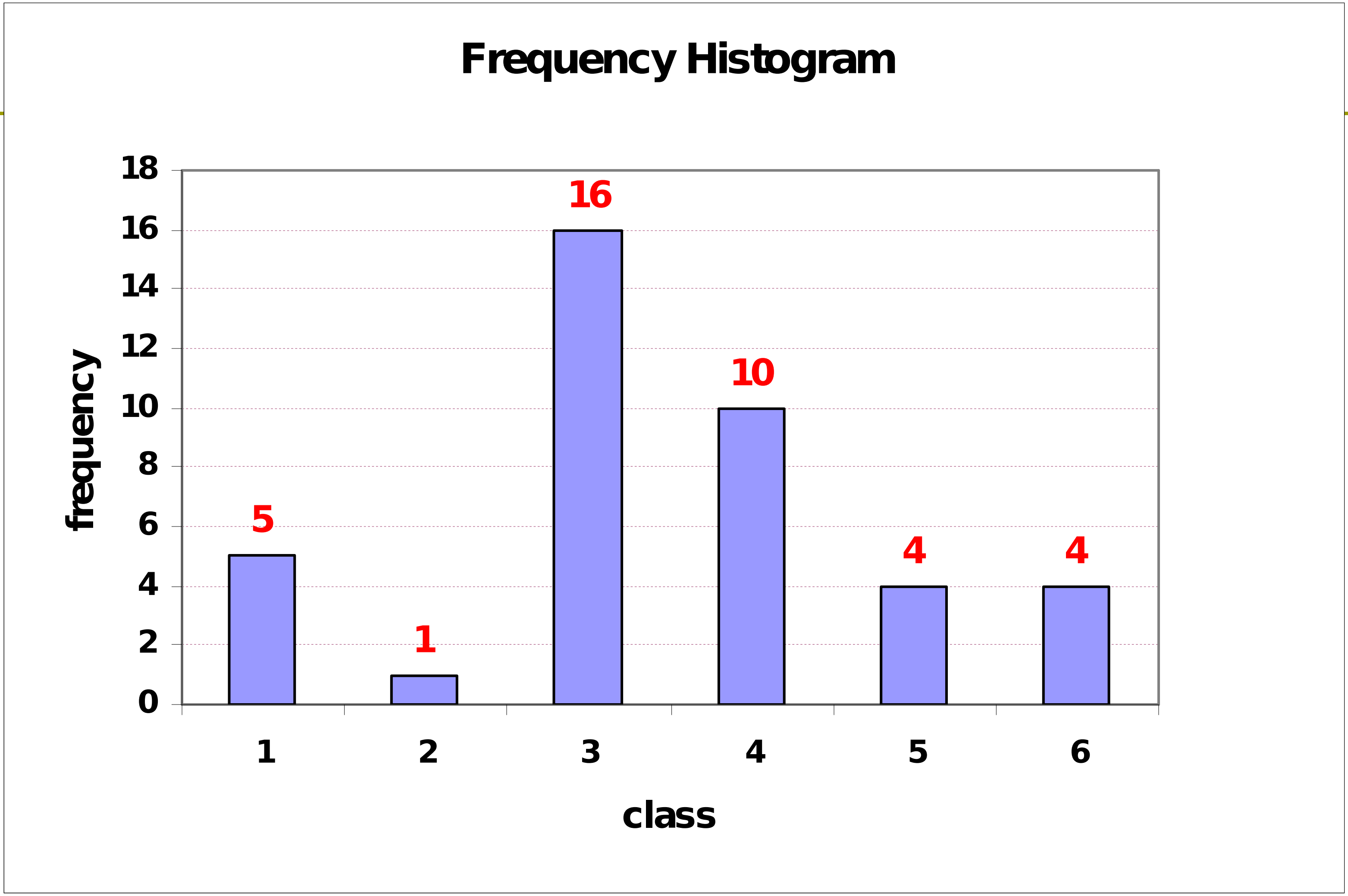
Consider the following data, construct the **frequency histogram**.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1.2 | 1.98 | 2.0 | 1.2 | 1.98 | 2.06 | 1.48 | 2.0 |
| 2.5 | **3.02** | 1.75 | 2.05 | 1.71 | 1.1 | 1.82 | 1.8 |
| 1.75 | 1.17 | 2.25 | 1.9 | 2.03 | 1.98 | 2.15 | 1.96 |
| 1.87 | **0.95** | 2.2 | 1.85 | 1.7 | 1.9[illegible] | 3.0 | 2.4 |

# Solution

- Number of classes $= \sqrt{40} = 6.32 \cong 6 \; classes$
- Class width $= \dfrac{3.02 - 0.95}{6} = 0.345 \cong 0.35$

| Class # | Class interval $0.95 \le x < 1.3$ | Frequency | Relative frequency |
|---|---|---|---|
| **1** | $1.3 \le x < 1.65$ | 5 | 0.125 |
| **2** | $1.65 \le x < 2.0$ | 1 | 0.025 |
| **3** | $2.0 \le x < 2.35$ | 16 | 0.4 |
| **4** | $2.35 \le x < 2.7$ | 10 | 0.25 |
| **5** | $2.7 \le x < 3.05$ | 4 | 0.1 |
| **6** | | 4 | 0.1 |
| | | **40** | **1** |


Frequency Histogram
18
16
14
12
10
8
6
4
2
0
frequency
16
10
5
4
4
1
1
2
3
4
5
6
class

# 6.2.4 Steam and leaf plot

The steam and leaf plot is similar to histogram, the data is split into a stem (the first digit) and leaf is the second digit.

| Stem (tens) | Leaves (units) | |
|---|---|---|
| 2 | 1244555566667777888899999 | |
| 3 | 0011122333334445555555677888899 | |
| 4 | 01111223569 | |
| 5 | 04 | ←Values are 50 and 54. |
| 6 | 013 | |
| 7 | 4 | |
| 8 | 0 | ←Value is 80. |

# Problem

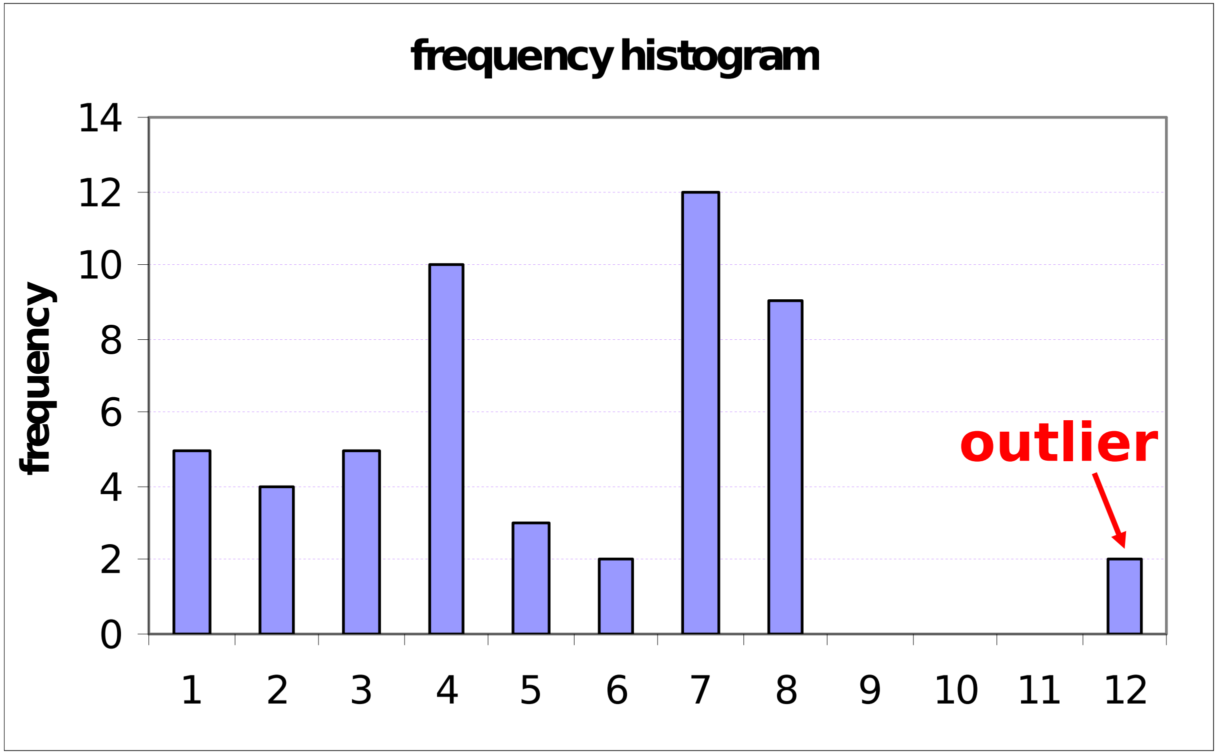
**For the Following set of data, using five classes Find:**

- The frequency distribution table
- The relative frequency distribution table
- The Frequency Histogram
- Present the data as stem and leaf plot

77 67 92 88 75 66 95 64 82 87
87 94 61 87 69 55 75 88 93 76 71
65 90 85 83 92 77 69 92 66 74
70 91 65 84 74

# 6.2.5 Outliers

- Outliers can be defined as the data points that appear to be separate from the rest of data.
- It is usually sensible to be removed from the data.



# Sample Quantiles

- **The sample median is the 50$^{th}$ percentile.**
- **The upper quartile is the 75$^{th}$ percentile.**
- **The lower quartile is the 25$^{th}$ percentile,**
- **The sample inter-quartile range denotes the difference between the**

# Example

- **Consider the following data:**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **0.9** | **1.3** | **1.8** | **2.5** | **2.6** | **2.8** | **3.6** | **4.0** | **4.1** | **4.2** |
| **4.3** | **4.3** | **4.6** | **4.6** | **4.6** | **4.7** | **4.8** | **4.9** | **4.9** | **5.0** |

**Find:**

- The sample media.
- The upper quartile.
- The lower quartile.
- The sample inter-quartile range .

# Solution:

1. <u>**The upper quartile:**</u>

**The 15th largest value = 4.6**

**The 16th largest value = 4.7**

**The 75th percentile** $= \left(\frac{1}{4}\times 4.6\right) + \left(\frac{3}{4}\times 4.7\right) = 4.675$

2. <u>**The lower quartile:**</u>

**The 5th largest value = 2.6**

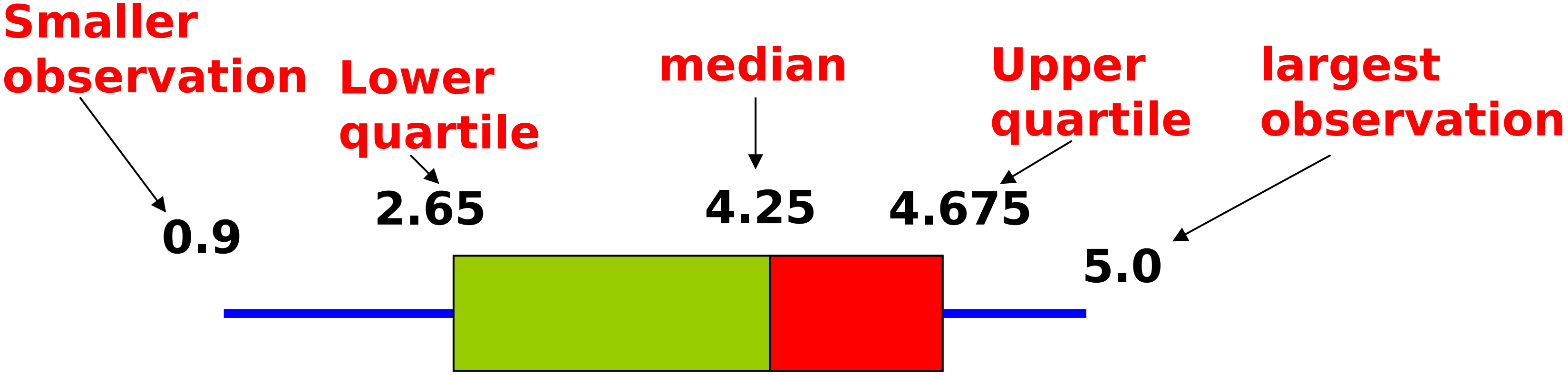
**The 6th largest value = 2.8**

**The 75th percentile** $= \left(\frac{3}{4}\times 2.6\right) + \left(\frac{1}{4}\times 2.8\right) = 2.65$

3. <u>**The inter-quartile range = 4.675 – 2.65**</u>

# Box Plots

- Pox plot is a schematic presentation of the sample median, the upper and lower sample quartile, the largest and smaller data observation.



# 6.3 Sample statistics

Provide numerical summary measures of data set:

- **Sample mean or arithmetic average ( $\bar{x}$ )**

$$\bar{x} = \frac{\sum_{i=1}^{n} x}{n}$$

# Median

**Is the value of the <u>middle</u> of the (ordered) data points.**

**<u>Notes:</u>**

**If data set of 31 observations, the sample median is the 16th largest data point.**

# Sample trimmed mean

- **Is obtained by deleting of the largest and some of the smallest data observations. Usually a 10% trimmed mean is employed ( top 10% and bottom 10%).**

  **For example, 10% trimmed of 50 data points, then the largest 5 and smallest 5 data observations are removed and the mean is taken for the remaining 40 data observations.**

# Sample mode

- Is used to denote the category or data value that contains the largest number of data observation or value with largest frequency.

$$s^2 = \frac{\sum (x - \bar{x})^2}{(n-1)}$$

- Sample variance:

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{(n-1)}}$$

- Standard Deviation:

# Example

**Consider the following data:**

**0.9 1.3 1.8 2.5 2.6 2.8 3.6 4.0 4.1 4.2**
**4.3 4.3 4.6 4.6 4.6 4.7 4.8 4.9 4.9 5.0**

**Find:**

- The sample mean
- The sample median.
- The 10 % trimmed mean.
- The sample mode.
- The sample variance.

# Solution

- **The sample mean:** $\bar{x} = \dfrac{\sum_{i=1}^{n} x}{n} = \dfrac{0.9+1.3+1.8+\dots\dots+4.9+5.0}{20} = 3.725$

- **The sample median** is the average of the 10th and 11th observation $= \dfrac{4.2+4.3}{2} = 4.25$

- **The 10 % trimmed mean:**
  Top 10% observation = 0.1*20 = 2 observations,
  and top 10% observation = 0.1*20 = 2 observations.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 0.9 | 1.3 | 1.8 | 2.5 | 2.6 | 2.8 | 3.6 | 4.0 | 4.1 | 4.2 |
| 4.3 | 4.3 | 4.6 | 4.6 | 4.6 | 4.7 | 4.8 | 4.9 | 4.9 | 5.0 |

$$trimmed\ mean = \frac{1.8+2.5+4.6+\dots\dots+4.8+4.9}{16} = 3.9$$

- **The sample mode = 4.6**

**0.9 1.3 1.8 2.5 2.6 2.8 3.6 4.0 4.1 4.2**

**4.3 4.3 4.6 4.6 4.6 4.6 4.7 4.8 4.9 4.9 5.0**

- **The sample variance.**

$$s^2 = \frac{\sum (x - \bar{x})^2}{(n-1)} = \frac{5009.101}{19} = 263.64$$

| x | x-x' | (x-x')$^2$ |
|---|---|---|
| 0.900 | -2.825 | 7.981 |
| 1.300 | -2.425 | 5.881 |
| 1.800 | -1.925 | 3.706 |
| 2.500 | -1.225 | 1.501 |
| 2.600 | -1.125 | 1.266 |
| 2.800 | -0.925 | 0.856 |
| 3.600 | -0.125 | 0.016 |
| 4.000 | 0.275 | 0.076 |
| 4.100 | 0.375 | 0.141 |
| 4.200 | 0.475 | 0.226 |
| 4.300 | 0.575 | 0.331 |
| 4.300 | 0.575 | 0.331 |
| 4.600 | 0.875 | 0.766 |
| 4.600 | 0.875 | 0.766 |
| 4.600 | 0.875 | 0.766 |
| 4.700 | 0.975 | 0.951 |
| 4.800 | 1.075 | 1.156 |
| 4.900 | 1.175 | 1.381 |
| 4.900 | 1.175 | 1.381 |
| 5.000 | 1.275 | 1.626 |
| **74.500** | | **5009.101** |